Thesis for the degree of Doctor of Engineering

Spatial analysis and modeling of nerve fiber patterns

Claes Andersson

CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
Chalmers University of Technology
and University of Gothenburg
Gothenburg, Sweden, 2018

This work has been financially supported by the Swedish research council (VR 2013-5212) and Wilhelm and Martina Lundgren ScienceFund.

Spatial analysis and modeling of nerve fiber patterns
Claes Andersson
ISBN 978-91-7597-735-5

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 4416
ISSN 0346-718X

Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg
SE–412 96 Gothenburg
Sweden
Phone: +46 (0)31-772 10 00
Author e-mail: andclae@chalmers.se

Cover: Four point patterns (left) and their pooled L-function along with a 95% pointwise confidence band obtained using a bootstrap technique (right). The same technique is used to compare groups in Paper I.

Typeset with LaTeX.
Printed in Gothenburg, Sweden, 2018.

Spatial analysis and modeling of nerve fiber patterns

Claes Andersson

Department of Mathematical Sciences
Chalmers University of Technology and University of Gothenburg

Abstract

Diabetic neuropathy is a condition associated with diabetes affecting the epidermal nerve fibers (ENFs). This thesis presents analysis methods and models for ENF data, with two main puroposes: to find early signs of diabetic neuropathy and to characterize how this condition changes the nerve fiber structure. Early detection is of interest to be able to take measures to slow down the progression of the condition, and a more detailed description of the changes in the nerve fiber structure could improve the understanding of its underlying mechanisms.

The ENF samples are mainly analyzed as point patterns, where the points are the locations where nerve fibers enter the epidermis or terminate. The analysis is partly based on existing summary statistics for point patterns, but we also propose a new summary statistic to quantify the proportion of the skin covered by the nerve fibers. Two cluster processes are introduced as models for the patterns consisting only of the locations where the nerve fibers enter the epidermis. For one of the models, a Bayesian hierarchical method for parameter estimation is proposed. A model for the end points is also presented, and non-spatial models for individual nerve fibers, which are used to perform unsupervised classification of the subjects.

From the results we find that while all patterns are aggregated, the level of aggregation tends to increase with increased severity of the neuropathy. The results from the modeling indicate that the increased aggregation is caused by a decrease in the number of clusters, while the structure within clusters appears to be similar in all disease groups. The results from the non-spatial analysis indicate that the nerve fibers from healthy subjects tend to extend further than those from subjects with diabetic neuropathy.

The use of methods and models developed in this thesis is not limited to ENF data, but can be applied to point pattern data in general. In particular, the models for the base point patterns and the methods for estimating the parameters of these models are contributions to the point process literature.

Keywords: Bayesian estimation, cluster processes, diabetic neuropathy, epidermal nerve fibers, hierarchical models, spatial point processes.

List of appended papers

The following papers are included in this thesis.

Paper I **Andersson, C.**, Guttorp, P. and Särkkä, A. (2016). Discovering early diabetic neuropathy from epidermal nerve fiber patterns. *Statistics in Medicine*, 35(24):4427-4442. doi: 10.1002/sim.7009

Paper II **Andersson, C.**, Rajala, T. and Särkkä, A. (2018). A Bayesian hierarchical point process model for epidermal nerve fiber patterns. *(Submitted to Spatial Statistics)*

Paper III **Andersson, C.** and Mrkvička, T. (2018). Inference for cluster point processes with over- or under-dispersed cluster sizes. *(Preprint)*

Paper IV **Andersson, C.**, Rajala, T. and Särkkä, A. (2017). Hierarchical models for epidermal nerve fiber data. *Statistics in Medicine*, 37(3):357-374. doi: 10.1002/sim.7516

My contribution to the appended papers:

Paper I I carried out the statistical analyses, and co-developed the concept of reactive territories. I co-developed the NOC model, and developed the estimation method. I did most of the writing for the publication.

Paper II I co-developed the model, the estimation method and the code for fitting the model. I did the statistical analyses, except for the posterior pair-correlation function envelopes and the IWE/ISWE-results. I did most of the writing for publication.

Paper III I co-developed the model, and did most of the work on developing the parameter estimation methods. I conducted the simulation study and did most of the writing for publication.

Paper IV I co-developed the models and the method for fitting the models. I did the statistical analyses and most of the writing for publication.

Acknowledgements

First and foremost, I want to thank my supervisor Aila Särkkä, for introducing me to the subject and always finding the time for meeting with me and discussing my work. Your support has helped me take on the many challenges I have faced during my doctoral studies. I would also like to thank my assistant supervisor Tuomas Rajala, for his ideas, suggestions and attention to details. You have truly helped me improve the quality of my work. I also want to thank Tomáš Mrkvička, co-author of Paper III. I really enjoyed working together, and my stay in České Budějovice was a timely boost for me. Also, many thanks to Peter Guttorp, coauthor of Paper I and an inspiration to work with, and Gwen Wendelschafer Crabb, William Kennedy and the others at Kennedy Lab, for answering questions about the data and neurology.

A big thank you to my colleagues at the department, both current and former. Ivar, for being a such good friend and someone to talk to through highs and lows, Anna R for our talks with a perspective going far beyond work, Malin for insisting on decent lunches, Henrike for being an excellent travel companion, and Mariana for sharing many fun moments. Also to Magnus, Peter, Dawan, Anna J, Fanny, Sandra, Tobias, and many more that have made this department such a great place to work.

Last, but definitely not least, a big thank you to my family for always being there for me, and for all their love and support over the years.

Contents

Chapter One

Introduction

Epidermal nerve fibers (ENFs) are thin sensory nerve fibers in the epidermis, the outermost layer of the skin. The function of the ENFs is to transmit signals of heat and pain, among other things. The ENFs enter the epidermis as single nerve fibers cross the junction between the epidermis and the dermis (the layer below the epidermis), and extend into the epidermis, with or without branching, before terminating. The existence of ENFs was theorized for over 130 years, before it was conclusively established by William Kennedy and Gwen Wendelschafer-Crabb, using confocal microscopy (Kennedy and Wendelschafer-Crabb, 1993). Figure 2.1 shows a side section of a skin blister with the nerve fibers visible (left panel). In the work presented here, the data are mainly analyzed as point patterns. The points are the locations where the nerves enter the epidermis, branch or terminate. These are henceforth referred to as base, branching and end points, respectively.

Once the methods for visualizing and identifying ENFs were established, research moved towards quantification of such nerve fibers, in particular to assess their potential use in diagnosing peripheral neuropathy, which is damage or disease affecting the nerves. The symptoms include loss of sensation and neuropathic pain, which can drastically reduce the life quality of an affected subject. In this thesis, we analyze data from healthy subjects, and subjects suffering from diabetic neuropathy, i.e. neuropathy caused by diabetes. Although there is no treatment to cure diabetic neuropathy, measures can be taken to slow down the progression. Therefore, it is of interest to be able to detect it as early as possible.

It is well established that neuropathy causes the number of nerves to decrease, and the use of ENF data in current clinical practice is largely based on ENF density, i.e. the observed number of nerves in the epidermis (Lauria et al., 2010b). A large number of studies have been dedicated to establish a normative reference range for the ENF density (see e.g. Lauria et al., 2010a). However, it has also been noted that the nerve fibers in subjects with diabetic neuropathy exhibit more clustered pattern than in healthy subjects (Kennedy et al., 1999). This observation has been quantified in several studies, using methods from spatial statistics. Using samples from the thigh of one healthy subject, one subject diagnosed with mild, one with moderate and one with severe dia-

betic neuropathy, Waller et al. (2011) found that the patterns from the subjects with moderate and severe diabetic neuropathy were significantly more clustered than the patterns from the healthy subject, in terms of second order summary statistics. Two papers have investigated the effect of three non-spatial covariates (age, gender and BMI) on the second order structure of the nerve fiber patterns. (Myllymäki et al., 2012) used a linear mixed model approach to investigate the effect on the second order summary statistic of the base point patterns from a group of healthy subjects. The conclusion was that the covariates affected the spatial structure in samples from the calf, but no significant effects of the covariates were found for samples from the foot. Furthermore, Myllymäki et al. (2014) included disease status (healthy or mild/moderate) as a covariate in a Gaussian process regression to model the second order summary statistic for samples from the calf. This study found no significant effect of any non-spatial covariate (not even disease status) on the spatial structure of the base point patterns, but found that the end point patterns were more clustered in the group of subjects with mild or moderate diabetic neuropathy. Also, Olsbo et al. (2013) proposed a point process model for the base and end point patterns of healthy volunteers, the non orphan cluster (NOC) model.

The work behind this thesis had two main aims: to find methods to detect diabetic neuropathy at an early stage and to describe the changes caused by diabetic neuropathy to the structure of the nerve fiber patterns in more detail. Early detection is of interest to be able to impede the development of the neuropathy and thereby delaying the symptoms, while the diagnostic techniques used today are often based on looking for such symptoms, e.g. by checking for loss of sensation. An improved description of the changes in the nerve fiber pattern, on the other hand, could give a better understanding of the underlying mechanisms of developing diabetic neuropathy. To our knowledge, not much is known about how diabetic neuropathy affects the nerve fiber growth, other than that the ENF density decreases. While these two aims often overlap, the papers in this thesis are often focused more towards one of them.

This thesis is in part a continuation of the work above, as tools from spatial analysis are a big part of the work, but to a large extent the work consists of developing new methods and models. A new tool to quantify the proportion of the skin covered by the nerve fibers is proposed. In addition, two new cluster point process models for the base point patterns and one for the end point patterns are introduced. Moreover, parameter estimation methods for these models are developed and evaluated. Finally, one paper takes a non-spatial approach, focusing on the characteristics of individual nerve fibers, which has not been under much consideration before. Here, models for the segments of the nerve fibers are developed, which are then used to perform unsupervised classification of the subjects.

The rest of the thesis is organized as follows. Chapter 2 describes the ENF

data set that is analyzed, at least partly, in all the appended papers. Chapter 3 presents basic definitions for spatial point processes, summary statistics and point process models. Thereafter, Chapter 4 describes the hierarchical structure of the data as well as the hierarchical modeling framework used in Paper II. A brief summary of the papers is found in Chapter 5. Conclusions and future work are presented in Chapter 6, which is followed by the appended papers.

Chapter Two

Data

The ENF data set is rich and has several rare features. The structure of the data is in itself unique, with groups, subjects, body parts, blisters and samples forming a hierarchy. Moreover, the end points form clustered patterns, where both the locations of parent points and the connections between parents and offspring are known, which is rarely seen. To be able to answer relevant neurological questions, these aspects of the data should ideally be taken into account, which often calls for developing new methods. Much of the work in this thesis is of that nature.

The data set consists of skin samples taken from 32 healthy volunteers and 20 diabetic subjects. The samples are obtained using suction induced skin biopsies, where a portion of the epidermis is removed, mounted on a slide and stained for imaging. The nerves are then traced using confocal microscopy, and the locations of the points where the nerves enter the epidermis, branch and terminate are recorded. The data are in 3D, and the samples are in boxes of size $330\mu m \times 432\mu m \times z$, where z varies from 20 to 50 μm. While the boundaries of the samples boxes are known in the x- and y-coordinates, there are no given boundaries in the z-coordinate. Although the base points appear at the bottom of the epidermis, i.e. the lower edge of the sample in the z-coordinate, the skin layer is not completely flat, which reveals itself through the varying z-coordinates of the base points. For the upper boundary the end points give a lower bound, but the end points do not necessarily appear at the upper surface of the epidermis. These difficulties, along with the fact that the main interest is the coverage of the area of the skin, led us to mainly restrict our attention to the 2D-projections in the x- and y-coordinates of the data. Illustrations of the types of data used in this work can be seen in Figure 2.1. In the top right panel ENF data are presented as a point pattern in 2D, where base points and end points are shown, and in the bottom right panel, part of the same sample with the branching points added can be found.

The data set consists of blister specimens taken from the right foot and right calf of each subject. Since the effects of the neuropathy tend to appear the earliest in the most distal parts of the body, we mainly focus on the samples taken from the foot. From each subject, three to six images were taken, typically two from each blister. Moreover, covariates such as age, gender and BMI were

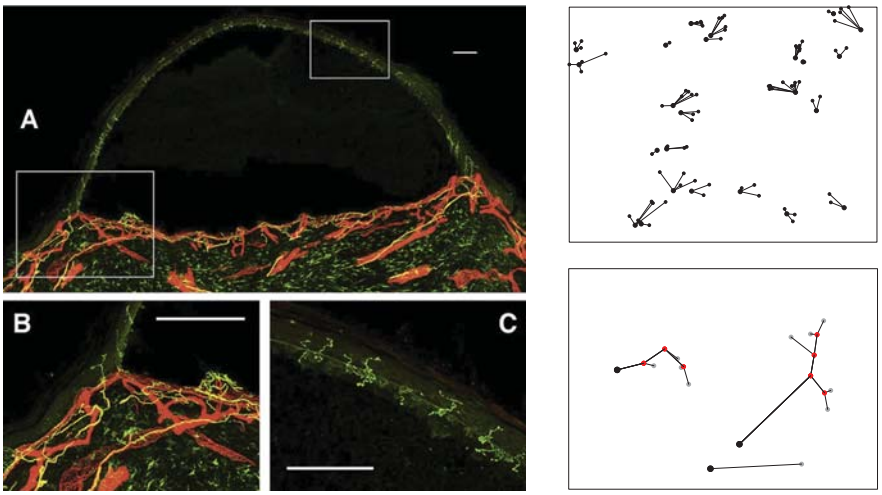

Figure 2.1: Left panel (source: Waller et al. (2011)): A side view of a suction induced skin blister, with the ENFs visible. The nerve fibers are from bright green to yellow, while blood vessels are red. Samples are collected as parts of the blister roof seen in Figure A. Figure B shows a dermal papilla, i.e. a ridge in the dermis, typically with an aggregation of nerve fibers. In Figure C the ENFs are zoomed in, with the nerve fibers clearly visible. Top right panel: ENF data illustrated as a point pattern. Large dots are base points and smaller dots are end points. Bottom right panel: part of the same sample with the base points as black dots, branching points as red dots and end points as gray dots.

recorded for each subject, as well as a neuropathy score (with higher score indicating more severe neuropathy). Moreover, base point densities in skin biopsies from six body locations were recorded for all diabetic subjects and part of the healthy volunteers. Based on the neuropathy score and the base point densities, the diabetic subjects were diagnosed as suffering from mild (little or no sign of neuropathy), moderate or severe diabetic neuropathy.

In Paper I the 2D projections of the x- and y-coordinates are used, and the data are treated as point patterns considering only the base points and end points. In Paper II and Paper III only the 2D projections of the base points are considered. Paper IV analyzes the 3D data using a non-spatial approach. Here, the segments connecting the base, branching and end points in each nerve are used as observations.

Chapter Three

Methods and models for spatial point processes

As mentioned earlier, one way to analyze the ENF data is to treat the data as a collection of point patterns, where only the information about the coordinates where nerve fibers enter the epidermis and where they terminate is used. When analyzing point patterns, a commonly occurring question is whether the points can be considered completely spatially random or if the underlying mechanism generating the patterns gives some structure to them. In Figure 3.1 three point patterns are visualized. The middle pattern is a realization of the case when the points are independently and uniformly scattered in the observation window, commonly referred to as complete spatial randomness (CSR). This means that the information that there is a point of the process at a certain location does not effect the probability of finding a point of the process anywhere else (in the observation window). Such a process is referred to as a homogeneous Poisson process, which can be defined through two properties. Firstly, there is a constant $\lambda > 0$ such that for any bounded set, B, the number of points in B follows a Poisson distribution with parameter $\lambda|B|$, where $|B|$ is the size of B, i.e. the area in the 2D setting. Secondly, given that the number of points in B is n, the n points are a random sample from the uniform distribution on B. The leftmost pattern, on the other hand, exhibits *regularity*. This means that the points of the pattern have a tendency to be further apart, compared to CSR. In other words, given the location of a point in the pattern, the probability of finding another point in its vicinity is lower than under CSR. The rightmost pattern of Figure 3.1 is an example of the opposite, where the points tend to appear in groups, which is called *clustering*.

The concepts of CSR, regularity and clustering are often used to characterize observed point patterns and the underlying process generating the patterns. In the following a brief introduction to the theory of point processes is given, leading up to a description of a few of the methods that can be used to analyze point patterns.

3.1 First definitions

Point processes have been used to study a wide variety of phenomena, including the locations of galaxies, trees in forest stands, ants' nests – and nerve fibers.

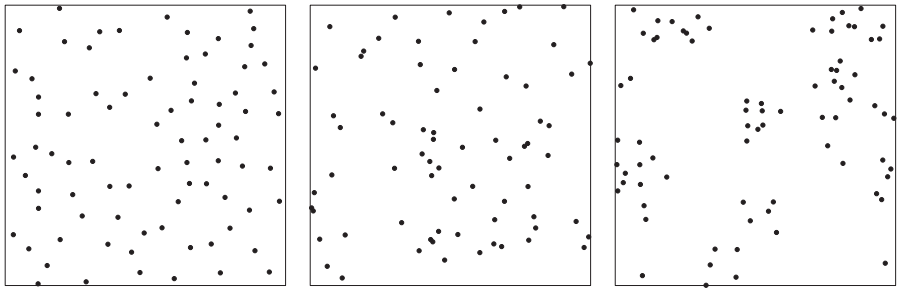

Figure 3.1: Realizations of three different point processes. The leftmost pattern is a realization of a process that exhibits regularity, the pattern in the center is a realization of a completely spatially random (CSR) process while the rightmost pattern is from a cluster process.

The plots in Figure 3.1 show realizations of point processes, i.e. the patterns are to the respective point process what a single number is to a random variable describing the roll of a die or what "heads" or "tails" is to a coin flip. The realizations of a point process are clearly more complex than those of the random variables mentioned above, and so is the theory to define and characterize point processes, as one might expect. Only a fraction of the theory will be brought up here, and described with more focus on intuition than mathematical rigour. More extensive and rigorous treatments of the subject can be found in e.g. Illian et al. (2008); Møller and Waagepetersen (2003); Chiu et al. (2013); Diggle (2014); Lieshout (2000); Gelfand et al. (2010), and Cressie (1993). The sections below mainly follow Illian et al. (2008).

Each of the patterns in Figure 3.1 can mathematically be described as an unordered set of points, $\mathbf{x}$, i.e

$$\mathbf{x} = \{x_1, \ldots, x_n\}, \tag{3.1}$$

where x_i describes the location of an observed point. Such an observation, $\mathbf{x}$, is commonly referred to as a *configuration*. Thus, a point process is a stochastic

mechanism, the realizations of which are random configurations of points in some space. Point processes can be defined on more general topological spaces, but here the attention will be restricted $\mathbb{R}^2$. Thus, a point process, X, in this setting may be defined as a random set of points

$$X = \{X_1, X_2, \ldots\} \tag{3.2}$$

where the number of points can be random or deterministic and where the locations, X_i, are random points in $\mathbb{R}^2$. Another way to define a point process is to view it as a random counting measure, that is, for any set $B \subset \mathbb{R}^2$ the point process maps the set to the random number of points of X falling in B. Here, this is used as a way to characterize the process, and $N_X(B)$ is the notation for the random number of points falling in the set B. A point process is *simple* if it places at most one point at any location and *locally finite* if for any bounded set B, the random variable $N_X(B)$ is finite.

Through the random variables $N_X(B)$ the *intensity measure* of a point process can be defined as $\Lambda(B) = \mathbb{E}(N_X(B))$, i.e. $\Lambda(B)$ is the expected number of points in B. Under some continuity conditions, which are usually satisfied in practical applications, Λ admits a density with respect to the Lebesgue measure, that is

$$\Lambda(B) = \int_B \lambda(x)dx, \tag{3.3}$$

where $\lambda(x)$ is the *intensity function* of the point process.

A point process is *stationary* if its distribution is translation invariant, and *isotropic* if its distribution is invariant under rotations around the origin. Note that for a point process to be stationary, it must be defined on the whole of $\mathbb{R}^2$. In practice, however, the observations of a point process are confined to some bounded region, which is only a part of the area on which the process operates. This is the case, for example, when observing a tree stand that is part of a larger forest or, as in the data used in this work, where the samples are very small compared to the whole skin. Although the full area of the process is also most commonly bounded, as is the case in the examples above, the point process is usually defined on $\mathbb{R}^2$. The underlying assumption is that the distance from the boundary of the full area of the process to the observation region is large enough for any boundary effects to be ignored in the observed region. One important property of a stationary process is that its intensity function is constant, i.e. $\lambda(x) \equiv \lambda$ for all x, which implies that $\mathbb{E}(N_X(B)) = \lambda|B|$, where $|B|$ is the area of B.

So far point processes have been described as models for the locations of points, but one could include more information by attaching *marks* to the points. Using the tree stand example, the marks of the points could be the diameter of the trees. For the nerve fiber data marks indicating whether a point is a base point or an end point are used. The class of marked point process models is

rich, as is the literature on the topic. Some of the earlier work can be found in Ogata and Tanemura (1985) and Takacs and Fiksel (1986), while Diggle (2014) and Illian et al. (2008) provide more recent examples. Since the use of marked point processes in this thesis is rather limited, the subject will not be treated further here.

3.2 Summary statistics

When analyzing point patterns, the aim is typically to reach an understanding of the process that generated the observed patterns. In this section some of the functional summary statistics that can be used for this purpose are introduced. They are all tools to describe the structure of the process, i.e. if the process generates clustered, regular or CSR patterns, or a combination of these. It will be assumed that the process considered is stationary and isotropic, although there are versions of some of the summary statistics for non-stationary and anisotropic processes as well. It should be noted, however, that unless some information that can explain differences in the intensity is available, there is no obvious way to differentiate between non-stationarity and clustering. The way in which the summary statistics are used is typically that they are estimated from data and compared to estimates from simulations under CSR or under some other null model. Some aspects of estimating these summary statistics from data are treated in Section 3.3.

One way to describe the structure of a point process is through the *empty space function*, $F(r)$, which gives the probability that the distance from an arbitrary location, x, to the nearest point of the process is less than or equal to r. As the process is assumed to be stationary, the random location can be replaced by the origin. Thus, this can be formulated as

$$F(r) = 1 - P(N_X(b(o,r)) = 0), \tag{3.4}$$

where $b(o,r)$ is a disc of radius r centered at the origin.

Replacing the location x in $F(r)$ by a point of the process, another summary statistic is obtained. This is called the *nearest neighbour distance distribution function*, and denoted by $G(r)$. The interpretation is that $G(r)$ is the probability that the distance from a randomly chosen point of the process to its nearest neighbour is less than or equal to r. Again, from stationarity, one can condition on having a point of the process at the origin, and express $G(r)$ as

$$G(r) = 1 - P_o(N_X(b(o,r) \setminus \{o\}) = 0), \tag{3.5}$$

where $P_o(\cdot)$ is the conditional probability given that there is a point of the process at the origin. Note that the probability of the process having a point at the origin is zero, so the conditional probability cannot be defined in the

classical way, by dividing by the probability of the event that is conditioned on. To properly define these conditional probabilities, one needs the concept of *Palm distributions*. A full introduction of this concept is outside the scope of this thesis, and the interested reader is referred to e.g. Lieshout (2000).

Under CSR, the points of the process are iid, which implies that $F(r) = G(r)$. For a clustered process, the distances between points of the process tend to be smaller than distances from an arbitrary location to a point of the process, which implies that $G(r) > F(r)$, while under regularity $F(r) > G(r)$. These relations indicate that it might be interesting to combine $G(r)$ and $F(r)$ in one summary statistic. The J-function (Lieshout and Baddeley, 1996) is such a combination, defined by

$$J(r) = \frac{1 - G(r)}{1 - F(r)}. \tag{3.6}$$

Under CSR, $J(r) = 1$, while $J(r) < 1$ indicates clustering and $J(r) > 1$ indicates regularity.

Another important summary statistic is Ripley's K-function, denoted by $K(r)$. The intuitive definition is that if X is a stationary and isotropic point process, with intensity λ, then $\lambda K(r)$ gives the expected number of further points within distance r from a typical point of the process. A strength of this summary statistic is that it contains information about the structure of the point pattern over a wide range of distances, while the previously mentioned $G(r)$ and $F(r)$ are more "short sighted". The mathematical definition of $K(r)$ is given by

$$\lambda K(r) = \mathbb{E}_o[N_X(b(o, r) \setminus \{o\})], \tag{3.7}$$

where the expectation is with respect to the Palm distribution mentioned above. It is also worth noting that the K-function is scaled by the intensity of the process. For example, if X is a CSR process, then $K(r) = \pi r^2$, which does not depend on λ. A related summary statistic is the *pair correlation function*, $g(r)$, which is given by

$$g(r) = \frac{K'(r)}{2\pi r}. \tag{3.8}$$

It can be noted that the relation to the K-function is very similar to that of a probability density function to the corresponding cumulative density function.

3.3 Edge corrections

The summary statistics mentioned above are defined in terms of a point process and when working with point pattern data one aim is typically to characterize the process that generated the data in terms of one or more of the summary statistics. Thus, estimators of the summary statistics are needed, but the naive estimators one would first think of are typically biased, due to edge effects.

Here, estimation of $K(r)$ will serve to illustrate some of the techniques available to compensate for edge effects.

As mentioned, $\lambda K(r)$ is the expected number of further points within distance r from a typical point of the process, and the common procedure to estimate an expectation from data is to replace it with the corresponding mean. An estimate of $K(r)$ can then be obtained as $\hat{K}(r) = \hat{\lambda}^{-1}\bar{n}(r)$, where $\bar{n}(r)$ is the estimate of the expectation $\lambda K(r)$ and $\hat{\lambda}$ is an estimate of the intensity of the process. Typically, $\hat{\lambda} = (n-1)/|W|$ is used, where n is the number of points in W and $|W|$ is the area of W, rather than the obvious choice $\hat{\lambda} = n/|W|$, for technical reasons. The mean, $\bar{n}(r)$, in this setting can be written as

$$\bar{n}(r) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}\mathbf{1}\{||x_i - x_j|| \leq r\}, \tag{3.9}$$

i.e. for each observed point the number of r-close neighbours is counted and the total count is divided by n. Unfortunately, this is a biased estimator of $\lambda K(r)$, and the reason for this is that the observed pattern is restricted to the window, W. It might be that a point, x_i, has neighbours that are within distance r but outside the observation window, and therefore not included in the sum (3.9). This introduces bias to $\bar{n}(r)$ in (3.9) as an estimator $\lambda K(r)$. If no information about the pattern outside W is available, as with the data used in this thesis, a common way to obtain an unbiased estimator is to introduce weights for the terms in the double sum, i.e. letting

$$\bar{n}(r) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i} w(x_i, x_j)\mathbf{1}\{||x_i - x_j|| \leq r\}. \tag{3.10}$$

Depending on the nature of the underlying process, different choices of w can be used. If the underlying process is (assumed to be) stationary, the *translation* edge correction can be employed, where

$$w(x_i, x_j) = \frac{1}{|W_{x_i} \cap W_{x_j}|}, \tag{3.11}$$

where W_{x_i} is the observation window translated by x_i. In other words, the weight is inversely proportional to the area of the intersection of the observation window translated by x_i and x_j. If the process is also isotropic the *isotropic* edge correction can be used, where $1/w(x_i, x_j)$ is the proportion of the circumference of the circle centered at x_i and having radius $||x_i - x_j||$ contained in W. This can be expressed as

$$w(x_i, x_j) = \frac{2\pi||x_i - x_j||}{l(\partial b(x_i, ||x_i - x_j||) \cap W)}. \tag{3.12}$$

where l is the length of a curve and ∂ denotes the boundary of a set.

A third example of an edge correction method is the *border correction*. With this method, for a given r, only points that are further away from the boundary than r are considered as reference points, while all points are considered when counting the number of neighbours for a reference point. Specifically, let n_r denote the number of points in $X \cap W$ that are at least at distance r from the boundary of W, and denote these points by $\{x_i'\}_{i=1}^{n_r}$. Furthermore, let $n_i(r)$ denote the number of r-close neighbours of x_i' among all the points in W. The mean can then be expressed as

$$\bar{n}(r) = \frac{1}{n_r} \sum_{i=1}^{n_r} n_i(r), \tag{3.13}$$

and $\bar{n}(r)$ can be used to estimate $K(r)$, as above, by using $\hat{K}(r) = \hat{\lambda}^{-1}\bar{n}(r)$.

3.4 Models

In the literature on point process models, there are several different classes of models. One very broad distinction can be made between models where the structure is a result from direct interaction between the observed points, and models where the structure arises from a latent random object. Examples of the former are found in the class of Gibbs processes. These were first developed in physics as models for particles, and the interaction between the points describes the forces the particles exert on each other. Gibbs processes are particularly useful for regular patterns, in which case an energy function depending on the pairwise distances between points is used. However, there are examples of Gibbs processes with more complicated energy functions as well as for clustered patterns.

Examples of models where the structure arises from a latent random object are Cox processes, where the observed process is a Poisson process with a random, spatially varying, intensity. Another example are cluster processes, where the observed points are clusters formed around an unobserved pattern of cluster centers, or parent points. Both of these examples are models for aggregated patterns. There is also an overlap between the two types, since under some assumptions a cluster process is also a Cox process. Both the base and end point patterns in the ENF data are aggregated, and we have mainly used cluster processes as models for them. Therefore, remainder of this section will treat the class of cluster processes and its relation to Cox processes.

Cluster processes

We shall refer to a cluster process, X, as a process which is a superposition of processes, as $X = \cup_{c \in C} X_c$, where C is a point process on $\mathbb{R}^2$, referred to as the parent process, and $X_c, c \in C$ are conditionally independent given C. Thus, we can construct a particular cluster process by specifying the following three components:

1. A point process model for C.
2. The distribution for the number of offspring, N_c, for each $c \in C$.
3. The distribution of the locations of the offspring relative to their parent.

A widely used subclass of cluster processes is the class of Neyman-Scott processes, which owes its name to the authors of two papers where cluster processes were proposed as models for the locations of galaxies (Neyman and Scott, 1952, 1958). This class is most often taken to be the subclass of cluster processes for which C is a homogeneous Poisson process (although the definitions in the original papers also allows for an inhomogeneous Poisson process for C) and the N_c's and offspring locations relative to their parents are iid. This definition is also adopted here. First and second order properties of a Neyman-Scott process can be expressed analytically. Letting κ be the intensity of C, $\alpha = \mathbb{E}[N_c]$ and $p_k = P(N_c = k)$, the intensity, λ, is given by

$$\lambda = \kappa\alpha$$

and the K-function is given by

$$K(r) = \pi r^2 + \frac{1}{\kappa\alpha^2} \sum_{k=2}^{\infty} p_k k(k-1) H(r), \quad r \geq 0,$$

where $H(r)$ is the distribution function of the distance between two points of the same cluster. The forms of $G(r)$ and $F(r)$ can also be derived, but explicit expressions for these are often difficult to obtain (see e.g. Chiu et al., 2013), and a formula for the J-function of Neyman-Scott processes has been derived (Lieshout and Baddeley, 1996).

It can be shown that a Neyman-Scott process with Poisson distributed offspring counts, N_c, is also a Cox process. This means that conditional on C, X is an inhomogeneous Poisson process, more specifically with intensity function

$$\lambda(x|C) = \sum_{c \in C} \alpha f(x - c), \tag{3.14}$$

where f is a the bivariate pdf for the offspring locations relative to their parents. Generalizing this, by letting the mean parameter of the Poisson distribution for

the offspring count of each parent be a random variable, yields the class of shot-noise Cox processes (see e.g. Møller and Waagepetersen, 2003, Chapter 5.4). Generalizing further, by relaxing the Poisson restriction on the parent process and allowing for the offspring location distribution to vary randomly between clusters one obtains the class of generalized shot-noise Cox processes (Møller and Torrisi, 2005).

Neyman-Scott processes and SNCPs have been used as a model for a wide range of data, such as the locations of trees in forests (Tanaka et al., 2008), galaxies (Neyman and Scott, 1958) and whales Waagepetersen and Schweder (2006). However, the offspring counts have with very few exceptions been assumed to follow a Poisson distribution, or a mixture Poisson distribution as for the SNCPs. Although this is quite flexible, it is limited to distributions for which the variance is greater than or equal to the mean. In Paper III we introduce a new cluster process for which the variance of the offspring count distribution can be smaller than, equal to and greater than the mean, with the Poisson distribution as a special case. This extra flexibility comes at the price of more challenging parameter estimation, but allows us to investigate the cluster size distribution in more detail. When the model is fitted to base point patterns from the ENF data, the results indicate that the variance in the cluster sizes is clearly lower than the expected value.

Chapter Four

Hierarchical models

The data we study come with a natural hierarchy to it since we have observations on different levels of the data, with levels ranging from groups of subjects to nerve fiber segments. Thus we can summarize the data at these different levels. This structure of the data, and how it is used in the modeling, is explained in the sections below.

4.1 Hierarchy in the data

The top level of the hierarchy is the groups of subjects, at which we can mainly use averages as observations. The next level is the subject level, where from each subject we have data from two body parts, foot and calf, and from each body part several blisters are collected. At the subject level we have, for example, covariates, such as age and BMI as observations. The blister level is not considered in the modeling for reasons explained below. From each blister we have several samples, which for the spatial analysis form the lowest level in the hierarchy with the point patterns as the observations. When we analyze individual nerve fibers with a non-spatial approach, as in Paper IV, the number of nerve fibers in each sample is observed and the next level is the individual nerve fibers. For each nerve fiber, we use the sum of the segment lengths and maximum order of segments as observations, with the order of a segment being the number of times the fiber has branched before this segment. The lowest level in this analysis is the segment level, where we consider the lengths and orders as observations.

When modeling the data, it is possible to disregard some of the levels in the structure. For example, in our analyses so far, we have found very little indication that samples from the same blister of one subject are more similar than samples from different blisters of the same subject. Therefore, the blister level is not considered, but all sample-level observations from one subject are assumed to have the same distribution. This observation is important in itself, since it suggests that the nerve fiber pattern of a subject is quite constant on this spatial scale, leading us to consider all samples from one subject as replicated observations of the same process.

When analyzing the individual nerve fibers, we also impose a type of hier-

archy within a level specifying the models by giving conditional distributions where some of the responses depend on others. For example, the length of a nerve fiber segment is modelled to depend on, among other things, the order of that segment. This is expressed as a conditional distribution of the length, given the explanatory variables for that response, one of which is the order.

4.2 Bayesian hierarchical models

One aim with the modeling of the data is to obtain a more detailed characterization of the differences in the nerve fiber structure between disease groups. One way to approach this would be to fit a model to all the samples from each group, and compare the estimated parameters. However, the nerve regeneration process is not expected to be identical for all subjects within a group, which in terms of the model means that the parameter values vary between subjects. To get a more accurate description of the data, both at the subject and group levels, this variation should be taken into account.

One way to achieve this is to formulate the model in a Bayesian hierarchical framework. We consider here the model for one group of subjects, and focus specifically on the case where we assume a cluster process for the point patterns. We let X_{ij} denote sample j from subject i, $\mathbf{X}_i$ all the samples from subject i and $\mathbf{X}$ all the samples from the group. In this setting, X_{ij} is assumed to be a realization from the process with parameters $\theta_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{ip})$. Moreover, for each observed pattern there is an unobserved parent pattern, C_{ij}, which is a realization from the parent point process. The parameters of the parent point process are assumed to be included in θ_i. The θ_i's are in turn assumed to come from a distribution with some parameters ψ, which are specific for the group. For ψ, we specify a prior distribution with parameters that are common for all groups, formulated as $\psi \sim p(\psi)$. This can be succinctly formulated as

$$\begin{aligned} X_{ij}|\theta_i, C_{ij} &\sim p_X(X_{ij}|\theta_i, C_{ij}) \\ C_{ij}|\theta_i &\sim p_C(C_{ij}|\theta_i) \\ \theta_i|\psi &\sim p_\theta(\theta_i|\psi) \\ \psi &\sim p_\psi(\psi) \end{aligned}$$

The unknowns of the model are the group level parameters, ψ, the subject level parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_S)$ and the latent parent patterns $\mathbf{C}$, using the same index notation for the parent patterns as for the observed patterns, and were S is the number of subjects in the group at hand. In the Bayesian setting, estimates of these are based on the posterior distribution conditional on data. Letting $\mathbf{X}$ denote all the data from one group, the joint posterior can

be expressed as

$$p(\psi, \theta, \mathbf{C}|\mathbf{X}) \propto p(\psi) \prod_i \left[p(\theta_i|\psi) \prod_j p(C_{ij}|\theta_i) p(X_{ij}|\theta_i, C_{ij}) \right], \tag{4.1}$$

where the normalizing constant is $p(\mathbf{X})$. To compare the groups in terms of the model parameters, we are interested in the marginal posterior for ψ, which is given by

$$p(\psi|\mathbf{X}) = \int p(\psi, \theta, \mathbf{C}|\mathbf{X}) d\theta d\mathbf{C}, \tag{4.2}$$

and for individual subjects we are interested in the marginal posteriors for θ_i's, given by

$$p(\theta_i|\mathbf{X}) = \int p(\theta_i, \psi, \mathbf{C}_i|\mathbf{X}_i) d\psi d\mathbf{C}_i \tag{4.3}$$

The integrals in (4.2) and (4.3) are typically intractable, since the integral with respect to $\mathbf{C}$ is over the space of locally finite configurations. Thus, other techniques must be employed to estimate the posterior distributions. We have used MCMC sampling to obtain samples from the posterior distributions of ψ, θ and $\mathbf{C}$.

4.3 Mixture models

The diabetic subjects in our data are diagnosed as suffering from mild, moderate or severe diabetic neuropathy based on nerve counts at six body locations and a neuropathy score, and in the Bayesian hierarchical framework we use this grouping to compare the spatial structure in different groups. However, using some other observations might lead to a different grouping of the subjects, revealing other aspects of the neuropathy. Part of the work in this thesis is dedicated to finding a grouping of the subjects reflecting the severity of the disorder, based on the observed nerve fiber structures, and the nerve fiber density. For this, mixture models are used, in the framework of the hierarchical models described above. A mixture model is based on the assumption that the observations in the data come from different classes, for which the distributions of the observations may differ. A general mixture model can be described as follows: let y denote the response variable, Z the class membership and n the number of classes in the model. The density of y can then be expressed as

$$p(y) = \sum_{z=1}^{n} P(Z = z) p(y|Z = z), \tag{4.4}$$

where $p(y|Z = z)$ is the density of y for a sample from class z and $P(Z = z)$ is the probability that a sample is from class z.

One way to use this class of models is to do unsupervised classification, meaning that the class memberships are modelled as unobserved random variables, with Z_i being the class membership of observation i. Fitting the model provides estimates of the parameters, $\hat{\theta}$, but also estimates of the probabilities of the class memberships given the parameter estimates, $\hat{w}_{iz} = P(Z_i = z|y_i, \hat{\theta})$. The latter estimates can be used to group the data into n classes, by classifying the observation y_i into $z^* = \arg\max_z(P(Z_i = z|y_i, \hat{\theta}))$. Perhaps the most popular method for fitting this type of models is the expectation maximization (EM) algorithm (Dempster et al., 1977). First, initial values of the parameters and the probabilities of class memberships, w_{iz}, are given. The algorithm then proceeds iteratively by updating the parameter values given the w_{iz}'s, and then updates the w_{iz}'s given the current parameter values.

As a simple example of this, consider the simulated data from a multivariate normal mixture distribution with two components, given in Figure 4.1, having means and covariance matrices

$$\mu_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix} \qquad \Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0.75 \end{bmatrix}$$
$$\mu_2 = \begin{bmatrix} 2 \\ 0 \end{bmatrix} \qquad \Sigma_2 = \begin{bmatrix} 0.5 & -0.25 \\ -0.25 & 0.5 \end{bmatrix},$$

and proportions $P(Z = 1) = 0.4$ and $P(Z = 2) = 0.6$. This model is fitted to the data, and in Figure 4.2 is an illustration of the results showing the true class and fitted class of each data point. The model captures the two clusters well, and only a few data points are misclassified. However, this is an ideal case when the correct model with the correct number of classes is fitted to the data, while in practical applications, the appropriate distributions and number of classes to use are typically not known.

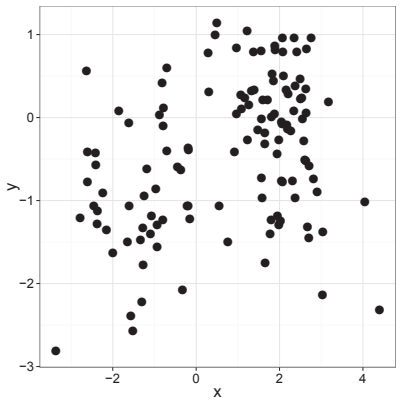

Figure 4.1: Simulated data from a multivariate normal mixture distribution.

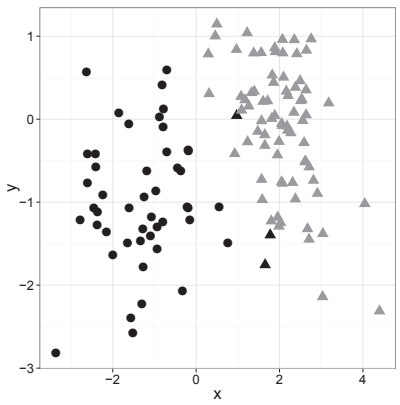

Figure 4.2: Illustration of the classification. The true classes are shown in different colors, and the fitted classes in different shapes. The black data points are from the class 1 and the gray data points from class 2, while the dots are classified as 1 and the triangles are classified as 2. Thus any gray circle or black triangle is a misclassification.

Chapter Five

Summary of papers

5.1 Paper I

In this paper a point process approach is used to compare samples from the right foot of 32 healthy volunteers to similar samples from 8 subjects diagnosed with mild diabetic neuropathy. The base point and end point patterns are treated as realizations of point processes and several summary statistics for point patterns are used to compare the groups. As part of the analysis we compare the pooled centered L-functions with bootstrap envelopes for the base point patterns from the two groups. The conclusion is that while the base point patterns are aggregated in both groups, they are more aggregated in the mild group than in the healthy group. Specifically, there is a range of distances for which the mean centered L-function takes larger values for the mild group than for the healthy group. The interpretation is that a larger proportion of the pairs of points are within these distances in the samples from the mild group than in the samples from the healthy group. The distances for which the centered L-function takes its largest value, which can be interpreted as a cluster radius, seems to be approximately the same in the two groups.

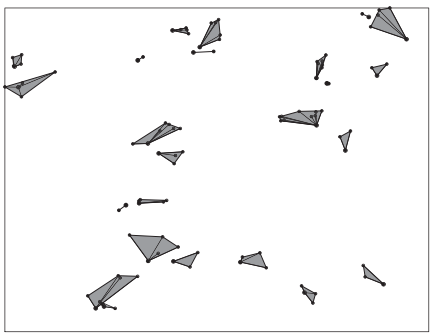

Figure 5.1: Illustration of the reactive territories in a sample. The dots are the base and end points, and the shaded areas are the reactive territories.

In the analysis we also use the information of which base point each end point is connected to, for example, to compare the distances between the base points and end points in the two groups. Moreover, a tool for quantifying the area of the skin covered by a nerve fiber is introduced, called the *reactive territory*. The reactive territory of a nerve fiber is defined as the convex hull of a base point and all its end points, as illustrated in Figure 5.1, and the coverage in a sample is the proportion of the observation window covered by reactive territories. In the comparison between the groups, it is found both that individual reactive territories are smaller and coverage is lower in the mild group than in the healthy group.

Finally, a point process model for the end points conditional on a base point pattern is introduced. The new model is compared to the non orphan cluster (NOC) model introduced by Olsbo et al. (2013). The modeling reveals that there are dependencies between the different nerve fibers of a more complex nature than what either model can capture. For the end points connected to the same base point, however, the new model seems to fit the data better than the NOC model.

5.2 Paper II

Here, a hierarchical model for the base point patterns is introduced. The model for the point patterns is the Thomas process, which is a cluster process. The interpretation is that the parent points are nerve trunks from deeper skin layers, and that base points from the same cluster are branches from the same nerve trunk. The model is constructed in a Bayesian hierarchical framework, meaning that there are subject specific parameters describing the patterns from each individual subject, while these parameters are assumed to be realizations of a distribution which depends on the disease group a subject belongs to. To estimate the parameters of the model we construct an MCMC algorithm. The algorithm samples from the joint posterior of the parameters and the unobserved parent patterns. The parent patterns are updated using a birth-death-move algorithm, while the subject-level parameters are updated using Metropolis-within-Gibbs sampling and the group parameters are updated using Gibbs sampling. The performance of the estimation method is evaluated in a simulation study, the results of which indicate that group level mean parameters are estimated quite well, but that the variation within a group is more challenging to estimate accurately. The model is fitted to the group of healthy subjects, the group diagnosed with mild diabetic neuropathy and a group of subjects diagnosed with moderate or severe diabetic neuropathy. By inspecting subject specific posterior pair-correlation functions we find that the model fits the data quite well, although there is a lot of variation between samples from the same subject. We also use integrated mean errors to get an overview of the model fit to all

subjects. From the posterior densities for the group level parameters, there is no indication of differences between the groups in the structure within the clusters. However, the posteriors for the parent point intensity indicate that there are fewer clusters per sample in the moderate/severe group than in the healthy group. The same posterior for the mild is located halfway between the posterior for the moderate/severe group and that of the healthy group, possibly indicating an early change in the structure of the base point patterns compared to the healthy group. However, there is substantial overlap between the parent point intensity posterior for the mild group and the same posterior for both other groups. Since a lower parent point intensity increases the clustering in terms of the L-function, this result gives a possible explanation for the differences found when comparing the groupwise L-functions in Paper I. The model fit also indicates that the number of points in each cluster is low, with an average well below 1, for all groups. This particular result led us into the work to further investigate the cluster size distribution.

5.3 Paper III

The results in Paper II motivated us to look for a model with a more flexible cluster size distribution, and in Paper III we introduce a new Neyman-Scott process that we call the Thomas process with generalized cluster sizes (TGCS). For a vast majority of cluster processes studied, including the Thomas process, the numbers of offspring per parent are Poisson distributed. For the TGCS, we instead let the cluster sizes follow a Generalized Poisson distribution (GPD). The Poisson distribution is a special case of the GPD, but the GPD has an extra parameter allowing the variance to be greater than or less than the expected value. The fact that the cluster sizes are not Poisson distributed means that many of the existing methods for parameter estimation for Neyman-Scott processes are insufficient for the TGCS.

As for many Neyman-Scott processes, the K-function and the pair-correlation function have closed form expressions for the TGCS. However, both of these functions are second-order properties, and they do not uniquely identify the TGCS. This means that two instances of the TGCS with different parameter values can have the same second-order properties. Thus, the minimum contrast method based on the K-function or pair correlation function is not enough to estimate the parameters of the TGCS. Instead, we introduce minimum contrast methods that use the K-function in a first step, and in a second step minimize the contrast for another summary statistic subject to constraints based on the results from the first step. As one option for the summary statistic to use in the second step, we introduce the so-called K^2-function, which is the expectation of the square of the further number of points within distance r from a point of the process.

We also construct a Bayesian MCMC algorithm for estimating the parameters of the TGCS. Since the TGCS is not a Cox process, this needs to take the parent-daughter connections into account in the sampling. This makes the updates of the parent pattern more intricate than in existing MCMC algorithms for Neyman-Scott processes. We also include a step where only the connections are updated. To evaluate the estimation methods we conduct a simulation study. We find that the MCMC algorithm works well, and in particular we correctly identify when the variance of the cluster sizes is less than, equal to or greater than the expected value in most of the cases considered. The results for the minimum contrast methods show that they are less accurate, especially using the K^2-function.

We fit the TGCS to four base point patterns from the ENF data. For comparison, we fit the Thomas process to the same patterns. We find that the cluster sizes seem to have a smaller variance than the expected value. The model fit indicates that the clusters are of size 0, 1 or 2. For all patterns, the probability of cluster size 0 is lower for the TGCS than for the Thomas process.

5.4 Paper IV

In this non-spatial analysis, the individual nerve fibers are in focus and are studied in terms of their tree structure which is composed of the segments between the points where the nerve starts, branches and ends. Summaries of the data both for the nerve fibers as a whole and for the individual line segments are introduced, and some initial exploratory analysis of the data from 32 healthy volunteers and 20 subjects diagnosed with mild, moderate or severe diabetic neuropathy indicate that there are some differences between the disease groups. The main question, however, is whether the subjects diagnosed with diabetic neuropathy can be separated from the healthy subjects based on such nerve tree data alone. For this, three models for the data are constructed, one for summarized information about each nerve fiber, one for information about individual segments and one for both types of information. All three models also use the number of nerve fibers in each segment. The models are formulated as mixture models, and fitted using the EM-algorithm. The mixture model approach allows us to use the models to perform unsupervised classification, and are evaluated using from 2 to 6 classes. Using the adjusted Rand index, which is a similarity measure of two partitions of a set, to compare the classification obtained by the model and the one based on the diagnosis, we conclude that the combined model with 4 classes is the best of those evaluated. It should be noted, however, that these 4 classes need not correspond to the 4 disease groups. The three models are compared to a baseline model that uses only the nerve counts from each subject, and the results indicate that the additional information on the tree structures clearly improves the classification.

Studying the grouping obtained from the fitted combined model with four classes, one class contains 11 out of 12 subjects diagnosed with moderate or severe diabetic neuropathy, 3 out of 8 of the subjects with mild diabetic neuropathy and 1 out of 32 of the healthy subjects. The remaining subjects are distributed over the other three classes. Looking into the properties of these classes, it seems that the class with the majority of the diabetic subjects is characterized by a low point count, short segments and few branching point before the nerve fibers terminate. In the remaining three classes, at least one of these features tends to be larger. One possible interpretation is that in subjects that are free of neuropathy, or where it has not developed very far, a decrease in one of these factors can be compensated by an increase in another. With this interpretation, this compensation does not occur in subjects with more severe diabetic neuropathy.

Chapter Six

Conclusions and future work

The work behind this thesis had two main aims: to explore methods for earlier detection of diabetic neuropathy and to improve the understanding of the changes in the nerve fiber structures caused by diabetic neuropathy. To achieve this, statistical analysis and modeling of ENF data were performed, mainly using and developing methods and models from spatial statistics. The use of this type of data in current clinical practice is focused on the base point density, but the results in this thesis indicate that also other types of information in these data could be useful for early detection. For example, in Paper I it was concluded that the distances between base and end points tend to be shorter in the subjects with mild diabetic neuropathy than in healthy subjects, and the results in Paper IV indicate that there are changes in the structure of individual nerve fibers in subjects suffering from diabetic neuropathy compared to healthy subjects. The work towards a diagnostic method using the type of data considered here is at an early stage, but the mixture model approach taken in Paper IV is a promising framework to build on. The models we considered could be expanded to include other types of information as well, such as scores from nerve conductivity tests.

As for the changes in the spatial structure in the nerve fiber patterns, the main focus has been on the base point patterns. The results in Paper II indicated that the base points appear in clustered patterns in all groups, but that the number of clusters decrease with the severity of the neuropathy. This result indicates that the base point density and the level of aggregation in the base point patterns are connected, and that these two changes could be explained by a change in a single aspect of the nerve fiber regeneration.

The methods developed here for the spatial analysis of the ENF data are not limited to this data alone, but can be applied to point pattern data in general. There are two main contributions to the spatial statistical methods. The first one is the modeling of data with a hierarchical structure, as in Paper II. While some methods have been developed for replicated data, i.e. where several point patterns are produced by an identical mechanism, there are few studies treating data with the nested structure that is analyzed in this thesis. Although the MCMC algorithm developed is specific for the Thomas process, it can easily be adapted to other cluster processes.

The second main contribution is the construction of a cluster process with a flexible cluster size distribution, and methods for estimating the parameters of such a model, as in Paper III. Again, although the estimation methods developed intended for estimating the parameters of the model introduced in the paper, they can be adapted to other cluster processes. To the best of our knowledge, a cluster process with this degree of flexibility in the cluster size distribution has not been studied before, and could be useful also in other applications.

Concerning possible future work, perhaps the most interesting path from a statistical point of view is the development of a joint 3D-model for the base and end points. In preliminary results, not included in this thesis, patterns consisting of the centroids of the end point clusters are less aggregated than the base point patterns. This indicates that end points tend to avoid end points connected to other base points. A very natural model for the end point patterns would be a cluster process where the parents are the base points, but such a model should also be able to capture interaction between the end point clusters. Currently, it seems that the cluster processes studied in the literature all build on the assumption of independent clusters. However, there are examples of spatio-temporal processes for earthquake occurrences that allow for this type of interaction. To adapt these to the ENF data, a way to incorporate the temporal aspect is needed.

Chapter Seven

References

Chiu, S. N., Stoyan, D., Kendall, W. S., and Mecke, J. (2013). *Stochastic geometry and its applications*. John Wiley & Sons: Chichester.

Cressie, N. A. (1993). *Statistics for spatial data*. Wiley Online Library.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B*, pages 1–38.

Diggle, P. J. (2014). *Statistical analysis of spatial and spatio-temporal point patterns*. Chapman & Hall: Boca Ranton.

Gelfand, A. E., Diggle, P., Fuentes, M., and Guttorp, P. (2010). *Handbook of spatial statistics*. CRC Press: Boca Ranton.

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical analysis and modelling of spatial point patterns*. John Wiley & Sons: Chichester.

Kennedy, W. R., Nolano, M., Wendelschafer-Crabb, G., Johnson, T. L., and Tamura, E. (1999). A skin blister method to study epidermal nerves in peripheral nerve disease. *Muscle & Nerve*, 22(3):360–371.

Kennedy, W. R. and Wendelschafer-Crabb, G. (1993). The innervation of human epidermis. *Journal of the neurological sciences*, 115(2):184–190.

Lauria, G., Bakkers, M., Schmitz, C., Lombardi, R., Penza, P., Devigili, G., Smith, A. G., Hsieh, S.-T., Mellgren, S. I., Umapathi, T., et al. (2010a). Intraepidermal nerve fiber density at the distal leg: a worldwide normative reference study. *Journal of the Peripheral Nervous System*, 15(3):202–207.

Lauria, G., Hsieh, S. T., Johansson, O., Kennedy, W. R., Leger, J. M., Mellgren, S. I., Nolano, M., Merkies, I. S., Polydefkis, M., Smith, A. G., et al. (2010b). European Federation of Neurological Societies/Peripheral Nerve Society Guideline on the use of skin biopsy in the diagnosis of small fiber neuropathy. Report of a joint task force of the European Federation of Neurological Societies and the Peripheral Nerve Society. *European Journal of Neurology*, 17(7):903–e49.

Lieshout, M. v. and Baddeley, A. (1996). A nonparametric measure of spatial interaction in point patterns. *Statistica Neerlandica*, 50(3):344–361.

Lieshout, v. M. (2000). *Markov point processes and their applications*. Imperial College Press: London.

Møller, J. and Torrisi, G. L. (2005). Generalised shot noise Cox processes. *Advances in Applied Probability*, 37(1):48–74.

Møller, J. and Waagepetersen, R. P. (2003). *Statistical inference and simulation for spatial point processes*. CRC Press: Boca Ranton.

Myllymäki, M., Panoutsopoulou, I., and Särkkä, A. (2012). Analysis of spatial structure of epidermal nerve entry point patterns based on replicated data. *Journal of Microscopy*, 247(3):228–239.

Myllymäki, M., Särkkä, A., and Vehtari, A. (2014). Hierarchical second-order analysis of replicated spatial point patterns with non-spatial covariates. *Spatial Statistics*, 8:104–121.

Neyman, J. and Scott, E. (1952). A theory of the spatial distribution of galaxies. *The Astrophysical Journal*, 116:144–163.

Neyman, J. and Scott, E. L. (1958). Statistical approach to problems of cosmology. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–43.

Ogata, Y. and Tanemura, M. (1985). Estimation of interaction potentials of marked spatial point patterns through the maximum likelihood method. *Biometrics*, 41(2):421–433.

Olsbo, V., Myllymäki, M., Waller, L. A., and Särkkä, A. (2013). Development and evaluation of spatial point process models for epidermal nerve fibers. *Mathematical Biosciences*, 243(2):178–189.

Takacs, R. and Fiksel, T. (1986). Interaction pair-potentials for a system of ant's nests. *Biometrical Journal*, 28(8):1007–1013.

Tanaka, U., Ogata, Y., and Stoyan, D. (2008). Parameter estimation and model selection for neyman-scott point processes. *Biometrical Journal*, 50(1):43–57.

Waagepetersen, R. and Schweder, T. (2006). Likelihood-based inference for clustered line transect data. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(3):264.

Waller, L. A., Särkkä, A., Olsbo, V., Myllymäki, M., Panoutsopoulou, I. G., Kennedy, W. R., and Wendelschafer-Crabb, G. (2011). Second-order spatial analysis of epidermal nerve fibers. *Statistics in Medicine*, 30(23):2827–2841.